MIT ESD

Massachusetts Institute of Technology
Engineering Systems Division

ESD Working Paper Series

A Unified Approach for Taxonomy-based Technology Forecasting

Andreas Henschel
Assistant Professor
Electrical Engineering and Computer Science
Masdar Institute of Science and Technology, UAE
Email: ahenschel@masdar.ac.ae

Erik Casagrande
Postdoctoral Researcher
Department of Chemical and Environmental Engineering
Masdar Institute of Science and Technology, UAE
Email: ecasagrande@masdar.ac.ae

Wei Lee Woon
Associate Professor
Electrical Engineering and Computer Science
Masdar Institute of Science and Technology, UAE
Email: wwoon@masdar.ac.ae

Isam Janajreh
Associate Professor
Department of Mechanical and Materials Engineering
Masdar Institute of Science and Technology, UAE
Email: wwoon@masdar.ac.ae

Stuart Madnick
John Norris Maguire Professor of Information Technology and Professor of Engineering Systems
MIT Sloan School of Management and MIT School of Engineering
Massachusetts Institute of Technology
Email: smadnick@mit.edu

ESD-WP-2014-29 October 2014 esd.mit.edu/wps

A Unified Approach for Taxonomy-based Technology Forecasting

Andreas Henschel
Erik Casagrande
Wei Lee Woon
Isam Janajreh
Stuart Madnick

Working Paper CISL# 2012-11

November 2012

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E62-422
Massachusetts Institute of Technology
Cambridge, MA 02142

Business Intelligence Applications and the Web:

Models, Systems, and Technologies

Marta E. Zorrilla
University of Cantabria, Spain

Jose-Norberto Mazón
University of Alicante, Spain

Óscar Ferrández
University of Alicante, Spain & University of Utah, USA

Irene Garrigós
University of Alicante, Spain

Florian Daniel
University of Trento, Italy

Juan Trujillo
University of Alicante, Spain

BUSINESS SCIENCE Reference

Senior Editorial Director:	Kristin Klinger
Director of Book Publications:	Julia Mosemann
Editorial Director:	Lindsay Johnston
Acquisitions Editor:	Erika Carter
Development Editor:	Myla Harty
Production Editor:	Sean Woznicki
Typesetters:	Chris Shearer, Milan Vracarich, Jr.
Print Coordinator:	Jamie Snavely
Cover Design:	Nick Newcomer

Published in the United States of America by
Business Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Business intelligence applications and the web: models, systems and technologies / Marta E. Zorrilla ... [et al.] editors.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-61350-038-5 (hardcover) -- ISBN 978-1-61350-039-2 (ebook) -- ISBN 978-1-61350-040-8 (print & perpetual access) 1. Business intelligence. 2. World Wide Web. I. Zorrilla, Marta E., 1971-
HD38.7.B8716 2011
658.4'7202854678--dc23
2011021833

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this book is new, previously-unpublished material. The views expressed in this book are those of the authors, but not necessarily of the publisher.

Chapter 8
A Unified Approach for Taxonomy-Based Technology Forecasting

Andreas Henschel
Masdar Institute of Science and Technology, UAE

Erik Casagrande
Masdar Institute of Science and Technology, UAE

Wei Lee Woon
Masdar Institute of Science and Technology, UAE

Isam Janajreh
Masdar Institute of Science and Technology, UAE

Stuart Madnick
Massachusetts Institute of Technology, USA

ABSTRACT

For decision makers and researchers working in a technical domain, understanding the state of their area of interest is of the highest importance. For this reason, we consider in this chapter, a novel framework for Web-based technology forecasting using bibliometrics (i.e. the analysis of information from trends and patterns of scientific publications). The proposed framework consists of a few conceptual stages based on a data acquisition process from bibliographic online repositories: extraction of domain-relevant keywords, the generation of taxonomy of the research field of interests and the development of early growth indicators which helps to find interesting technologies in their first phase of development. To provide a concrete application domain for developing and testing our tools, we conducted a case study in the field of renewable energy and in particular one of its subfields: Waste-to-Energy (W2E). The results on this particular research domain confirm the benefit of our approach.

DOI: 10.4018/978-1-61350-038-5.ch008

INTRODUCTION

Any given research field is composed of many subfields and underlying technologies which are related in intricate ways. A solid understanding of how these subfields are linked together as well as how important the different regions of this research landscape are will confer a significant competitive advantage. Currently, information regarding past and current research is available from a variety of channels, providing a rich source of data with which effective research strategies may be formed. These two important trends strongly motivate the development of computational tools for exploiting this data: firstly, the proliferation in technical and academic publications has greatly increased the rate at which relevant knowledge and data are produced and disseminated; secondly, access to this information is constantly improving thanks to the advances in the technologies underlying the web.

Motivation

In order to clarify the intended use of our system, it must be stressed that we are not using "forecasting" in the sense of weather forecasting, where future outcomes are predicted with a reasonably high degree of certainty. It is also important to note that certain tasks remain better suited to human experts. For example, where a particular technology of interest has already been chosen, we believe that a traditional literature review would prove superior to an automated approach. Instead, the proposed framework targets the preliminary stages of technology management, where breadth rather than depth is emphasized. The main focus of our system is on analyzing broad trends occurring in a very large number of documents or other textual sources. By scanning and digesting large amounts of information, promising but less obvious developments can be detected and subsequently brought to the attention of a human expert. This way we capitalize on the strength of computational approaches before making more efficient use of valuable expert time in the critical latter stages of the decision making process.

Knowledge that facilitates forecasting the likely growth and consequences of emergent technologies is essential for well-informed technology management, which is currently relying largely on expert opinion. However, expert decisions can be influenced by personal perspectives or biases. Moreover, acquiring and analyzing such knowledge is hampered by the vast amount of data available in publications. Consequently, sifting through the—often electronically—available R&D literature is time consuming, yet non-exhaustive and subjective. In order to cope with this problem, automated forecasting techniques have been developed in recent years (see Background section). A remaining challenge is related to the knowledge organization of the acquired data. For example, in order to elucidate the advances of technologies, we want to answer questions like: "How many scientific articles have been published in peer-reviewed journals on the topic of solar energy recently?" Intelligent search techniques capable of grouping semantically similar concepts are therefore needed, such that the term "parabolic trough" is subsumed under solar energy related technologies and hence articles about it should be included in the analysis. This underlying challenge of managing and structuring the vast amount of available knowledge from web resources is similar in web-based Technology Forecasting and general Semantic Web applications. However the former has yet to fully benefit from the advances of the latter. In particular, the state-of-the-art Technology Forecasting tools hardly make use of ontologies or taxonomies, the standard form of knowledge representation for the Semantic Web (Shadbolt, 2006).

The major novel aspect in the presented work is a modular and automated approach, which streamlines data acquisition, keyword extraction and taxonomy creation as the basis for trend detection. The framework provides evidence for growing technologies to decision makers in a logically structured manner. We therefore believe that it carries an enormous business potential within the realm of technology management.

BACKGROUND

Technology Forecasting. In order to assist purely expert-based technology management, automated bibliometrics techniques (i.e. the analysis of scientific publications) have been developed (Kostoff, 2001; Daim et al., 2006; Martino, 2003; Porter, 2005; Porter, 2007). These works perform bibliometric analyses in various ways: most commonly used are publication per year statistics for single keywords; other approaches deal with the inter-relationships between research topics, the identification of key authors and their collaboration patterns, the study of research performance and the core competences per country, institute or company.

The open problems with the existing approaches are related to the lack of structuring of the available input data: for example, when analyzing the trend of "Solar cells", the conventional methods will ignore documents on "Amorphous silicon", if they do not explicitly mention the main category. Also, within a research field, such as Photovoltaics, conventional approaches cannot detect the strongest contributing subfields (Organic Photovoltaics) because of the lacking concept hierarchy. Finally, they do not deal with vocabulary mismatch (synonyms, alternative wording).

Data acquisition and corpus generation. The main challenge of general data acquisition and data integration is that data has to be represented using the same abstraction principles. This challenge is frequently tackled with ontology-based approaches (Noy 2004, Kalfoglou 2003). They allow for the semantic integration of heterogeneous databases (i.e. the detection of correspondences in database schemata). In the case of web-based Technology Forecasting, we can draw from a number of diverse sources such as scientific publication databases, patent collections and blogs, however, the actual challenge of data integration boils down to the identification of entities that include a textual representation (abstracts or full text documents, blog entries, patent abstracts, claims), timestamps, and possibly keyword/tag annotations and authors including their affiliation/country. Ideally, the relevant, distributed databases would be accessible as a seamless, unified virtual data warehouse, interlinked through query interfaces such as SPARQL. Unfortunately, many database web front-ends are not configured for automated querying and data acquisition and access to large scale document and patent collections in machine-readable formats is rarely permitted to the public. Notable exceptions include the Open Archive Initiative (OAI) repositories machine. The OAI provides standards for web content interoperability such as a unified protocol for harvesting meta-data of publications. In the context of bibliometric analysis it is promising to see that as of 2009, 20% of all publications are now freely available, though not all in OAI archives (Björk, 2009). Blogs are commonly provided as RSS feeds but as a data source, it is important to keep in mind that they often contain only poorly structured information with the least amount of peer review.

Term Extraction. With the advent of the semantic web era, many communities and distributed companies simultaneously access and update textual online resources. For several web applications it is important to model the knowledge domain of such virtual communities. One of the fundamental steps of this modeling process is the modeling of the used vocabulary (i.e. the identification of domain-relevant terms). This step is often referred to as "keyword extraction" or "term extraction", in this context we also include keyphrase extraction, as English research terminology often includes compound word phrases.

The goal is to identify significant terms from a given corpus. The algorithms therein describe term extraction methods using statistical and linguistic features. In general, terms are detected after removal of stop words (words of little or no information gain such as "and", "the", "like"). The state of the art of automated techniques includes four different approaches:

- Techniques from Statistical Natural language processing. In order to detect potential candidates, N-gram (often unigram, bigram and trigram) models have been suggested (Manning, 1999). They provide a mean to estimate the probability of observing a phrase in a text using conditional probabilities of the n-1 preceding words. Further estimates for the significance of a term in a document are term frequency (TF), inverse document frequency (IDF), their product (TFIDF) and the position of a term in a document.
- Linguistic features for keyword extraction have been proposed, such as part of speech tags and part of speech tag patterns (for phrases), have been proposed (Hulth, 2003)
- Supervised Machine Learning techniques. They take as training data a set of documents for which keywords have been assigned manually. Documents are represented with features using the abovementioned techniques (Turney, 2000; Hulth, 2003).

It is important to emphasize that simple features as employed in KEA (Witten, 1999) using TFIDF and Naive Bayesian Classifiers perform reasonably well in comparison to sophisticated Machine Learning approaches. Moreover, statistical methods do not require any training data, are straightforward to implement and run fast. For a thorough review of keyword extraction methods, the reader is referred to (Pazienza et al., 2005).

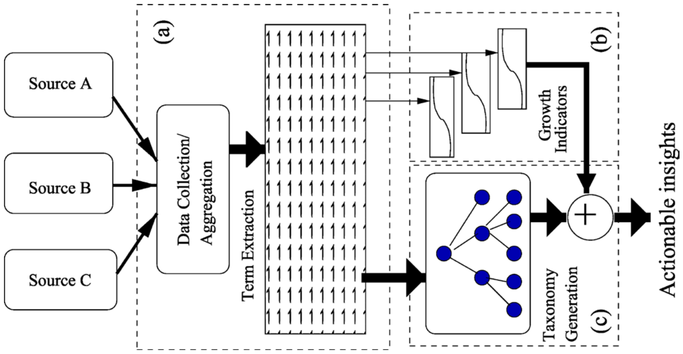
A number of online web tools for term extraction from text corpora exist, such as TerMine[1] (Frantzi, 2000) and in Yahoo!'s Query Language YQL[2]. In the case of the Yahoo! term extraction, a programmable interface (API) is provided. These web-based tools are useful as long as the demand of the user does not exceed the limitations, these web servers are commonly afflicted with. The limitations include volume restricted corpus size, the number of queries or restricted amount network traffic.

Taxonomy Creation. Taxonomies for scientific research bodies facilitate the organization of knowledge. They are applied in Information Retrieval tasks such as Semantic Web search engines (Shadbolt et al., 2006) and semantic database integration (Noy, 2004), where it is beneficial to abstract from plain words to hierarchical concepts. Many approaches for Ontology learning (Maedche, 2001) include the creation of taxonomies as a preliminary step. The commonly used techniques can be summarized as follows:

- Lexico-syntactic patterns (Hearst, 1992)
- Hyponymy information from WordNet (Miller, 1990)
- Noun phrase head matching (Navigli, 2003)
- Information theoretic approaches (Sanderson, 1999);
- Graph-theoretic approaches (Heymann, 2006)

The first approach is afflicted with a low recall in most corpora, whereas the latter approaches rely on the distributional hypothesis introduced by (Harris, 1968). It states that two words that appear in many similar linguistic contexts are semantically similar. However, this view is afflicted with the challenge that general terms such as "energy" and "fossil", or terms that somehow interact (e.g., "hammer" and "nail"), frequently co-occur and hence exhibit a misleadingly high co-occurrence similarity. Yet neither are subsumable in the strict sense ("is-a" or "part-of" relations) of standard taxonomies. Therefore, machine learning approaches have been used as meta-classifiers

Figure 1. The generic technology forecasting framework described in this chapter



(Cimiano, 2004). They combine several of the above mentioned techniques. Yet, the authors report that the best of these classifiers, a Support Vector Machine, only reached 33% percent F-measure. Likewise, a comparison of seven fully automated state-of-the-art taxonomy creation tools exhibited mediocre results with F-measures not exceeding 50% (Waechter, 2009). These results explain, why semi-automated methods are still frequently used.

The remainder of this chapter is organized as follows: The next section introduces a generic framework for a complete Technology Forecasting workflow. We discuss each module individually and how it is integrated in the framework. The usefulness of this methodology is demonstrated in the subsequent section, where we describe proof of concept implementations and their applications to case studies. Finally, we outline potential future research directions and draw conclusions in the last section.

FRAMEWORK

Over the past two years we have been working on a variety of taxonomy-based techniques for performing Web-based Technology Forecasting, (Woon and Madnick, 2009; Woon et al., 2009). While there are existing studies which address various aspects of this problem, it appears that an integrated and automated framework which can produce concrete, actionable results has not yet been satisfactorily presented. This chapter presents a detailed explanation of our approach, which goes some way towards addressing this need.

The high-level organization of our system is shown in Figure 1. To facilitate discussion, the system has been divided into the following conceptual blocks:

1. Data collection from web sources,
2. Term extraction,
3. Design of growth indicators,
4. Taxonomy generation,
5. Visualization of accumulated growth indicators.

These blocks and their inter-relations are shown respectively in Figure 1. In the following subsections these will be discussed in more detail. However, we would like to stress that the modular nature of this framework means that individual parts can be exchanged or modified without affecting the overall functionality of the system.

Data Acquisition and Corpus Generation

The initial process of data acquisition, shown in Figure 1(a), can draw from a broad variety of available sources. They differ in the breadth of their respective focus areas, accuracy, recency and machine readability. The corpus generation process unifies these heterogeneous data sources and ensures the integrity of important meta-data such as time stamps, document identifiers, authors and possibly keyword annotations and document titles/headers. In principle we consider blogs, conference papers, journal papers and patents (listed in order of increasing accuracy and decreasing recency). Data from the latter three is preferably acquired through professionally curate citation repositories such as Scopus, Google Scholar, Scirus, Compendex, PubMed, ISI Web of Science, the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), and the Derwent World Patents Index.

Further improvements could include author network analyses (see refs in the Future work section), which would impose additional requirements such as authors, their affiliations and possibly further information in order to disambiguate author names.

Term Extraction

Once a corpus of sufficient relevant documents is generated, we proceed with the task of Term Extraction (see Figure 1 (b)). Where applicable, i.e when documents are annotated already in scientific databases, keyword annotations can be used to guide the term extraction process: a simple approach would be rank these keywords according to their respective frequencies, where the most frequent keywords are collected and used as the domain vocabulary.

Multiple-word noun phrases are essential for Technology Forecasting, since many technology descriptions in the English language are composed of more than one word. Noun phrases can be detected with reasonable accuracy using a chain of state-of-the-art tools from Natural Language Processing, typically sentence and word tokenization, part-of-speech tagging and noun phrase chunking (Bird et al., 2009). Noun phrases can be compounds of nouns (“waste combustion”), adjective noun phrases (“thermal treatment”), prepositional and noun phrases (“board of directors”). Statistical significance of words or word phrases can then be estimated using information retrieval measures, in order to avoid irrelevant terms which are not specific for a certain technology (e.g. “review” or “approach”). These terms cause confusion in the downstream data processing such as the taxonomy creation. Ideally the extracted terms should be of low ambiguity and high specificity. These properties are preconditions for the conceptualization of a knowledge domain and the creation of a domain taxonomy/ontology.

One advantage of using an automated term extraction tool in addition to a manual one is that emerging technology terms and research field names can be potentially detected before they are recognized and consistently attached as keywords to articles. It also allows for working with data sources having poor or no keyword annotations, such as a variety of blogs.

Because of the aforementioned limitations in state-of-the-art web-tools we describe a small, powerful tool for offline usage in the next section. It provides full control over the selection of linguistic and statistical features in the term extraction process. Moreover, full control of the term extraction process is useful, if more customization towards a certain purpose is desirable (e.g. in the context of technology forecasting, it is easily possible in combination with the local literature database to extract only terms from a certain subset of the corpus, such as only recent documents).

Finally, it should be mentioned that corpus generation and term extraction can be viewed as an iterative, co-evolving process: newly extracted keywords can be used to extend the corpus, which in turn effects the next iteration of term extraction.

Growth Indicators

Given a set of keywords, we strive to find a suitable measure of their prevalence (Figure 1(c)). We hereby focus on keywords with relatively low but recently increasing occurrence frequency, which we refer to as "Early Growth" phase. As an easily derivable indicator in the context of bibliometric trend analysis it is helpful to look at the absolute amount of the growth rate of recent publications. A second step is to normalize the annual publication rates either by total volume of that particular technology or by total publication rate for all technologies. One particular growth indicator we focus on is the average publication year for a certain technology, defined in the following Equation 1.

$$\theta_i = \frac{\sum_{y} \in y \cdot TF_y[t_i]}{\sum_{y} \in yTF_y[t_i]} \qquad (1)$$

where θ_i is the growth potential for keyword t_i and $TF_y[t_i]$ is the term frequency for term t_i for a given year time span Y under consideration. This measure reflects the majority of publications, irrespective of its volume. Consequently, it facilitates the detection of recent topics, even small ones. Conclusions about trends with low publication volume must be drawn cautiously though, as they are prone to artifacts.

While curve fitting approaches are also commonly utilized, we refrain from using this approach for two main reasons: firstly, data might often follow an unexpected distribution that cannot be fitted using a preconceived shape. Renewable Energy technologies are influenced by political issues such as oil price regulations. As a consequence, it can be seen that "Solar thermal power" experienced a revival in recent years after being a hot topic in the early 80's (coinciding with the oil crisis of 1979) followed by a decrease of activity during the mid-80's until the mid-90's. For instance see (Dawelbait et al, 2010). The concept of technology revival or other turbulences is generally not reflected by trend discovery techniques such as the Fisher-Pry model and Gompertz (S-shaped) curves. Secondly, the statistical stability of a trend discovery relies on the exclusion of artifacts such as noisy term frequencies combined with low document coverage. As will be seen, our use of accumulated term statistics mitigates this problem to some extent. Yet, fluctuations are still present, especially when working with small corpora. For those cases, curve fitting methods would be equally inappropriate.

Taxonomy Generation

Taxonomies and ontologies (which additionally include non-taxonomic semantic relations for concepts) have been used for similar problems in the field of information retrieval (Wang, 2009). A feature of this approach is that a technology can be further analyzed in terms of its subcategories. If a general technology such as "Solar power" is on the rise, it is possible to retrieve a more differentiated view of the individual contributors thanks to the hierarchical nature of the used taxonomy. Unfortunately, in many cases a suitable taxonomy is not available. Moreover, manual taxonomy construction is costly and subjective.

In order to make our framework broadly applicable, we developed automatic and semi-automatic taxonomy creation algorithms (Figure 1(d)). In particular, in this study we consider a taxonomy creation process based on the Heymann-Algorithm (Heymann and Garcia-Molina, 2006). We previously considered as well two other approaches as described in (Woon and Madnick, 2009). The original aim of the Heymann-algorithm was the analysis of social tagging systems, where users collaboratively annotate a body of documents via the use of topical labels, also known as "tags". Inversion of this information results in a look-up table where each tag is associated with a vector that contains the frequencies of annotations for all documents. The Heymann-algorithm consists of the following two stages:

1. Firstly, a similarity scores are used to create a weighted graph of tags; this is then used to calculate the centrality of each of the tags. In graph theory, centrality is a measure of the connectedness of a node in a graph (a few approaches to calculating graph centrality exist (see our description in [Henschel et al., 2009] for example).
2. The tags are then ranked according to their respective centralities, and are inserted into a growing taxonomy in accordance to this ranking; the attachment of the tags is also determined by the similarity measure described above, where each tag is attached to either the most similar tag or to the taxonomy root.

Both parts involve similarity measures between terms. Originally the authors in (Heymann and Garcia-Molina, 2006) used vectors $\mathbf{x}_t$=[x_1,…,x_N] of length equal to the number of documents N, where x_1 describes, how many times a numbered document *i* in a user community has been annotated with term t. In Equation 2, we adapt this to binary term-vectors (or set representations) indicating whether a term occurs in a document (1) or not (0). Standard cosine vector similarity is therefore applicable.

$$S_{\cos}(x,y)=\frac{x\cdot y}{||\,x\,||||\,y\,||} \tag{2}$$

where **x** and **y** are binary term vectors. Hence, the similarity between two terms is simply the dot product of its normalized term vectors. We discuss several aspects that can be modified in order to boost the algorithm: generality ordering, similarity measures and weight functions insertion of new nodes. Other measures of distance are possible. For example, in (Woon and Madnick, 2009) we proposed an asymmetric distance function which is used to reflect the distances of terms that are in a taxonomic “is-a” relationship.

Complementary to fully automated taxonomy generation methods, we explore in this study the utilization of ontological background knowledge. This knowledge can be assumed to be previously indicated by an expert of the particular field of interest. Alternatively, such information may be extracted from online resources such as Wikipedia, a large-scale and accurate resource well-suited for Semantic Web and Information Extraction applications. (Giles, 2005, Auer et al., 2007). In both cases, the previous Heymann-algorithm is modified by starting its computation by a preexisted prior taxonomy. While the initial structure of such taxonomy maybe suboptimal by overconstraining the previous exploratory analysis and may hide interesting patterns in the data, it is possible that a semi automatic algorithm may be of a better interpretation by an expert in the field. This may help to refine his knowledge and modify in case his prior taxonomy. In the following case study section, we discuss both automatic and semiautomatic generation modality.

Growth Indicator Accumulation and Visualization

Finally, with the keyword taxonomies we can recalculate the early growth indicators based on aggregate scores of each of the individual scores of subordinate keywords contained in the according subtree of the taxonomy (Figure 1(e)). Several ways of aggregation are possible (i.e. the growth indication scores of subordinate terms can be either weighted equally or, for example, in terms their associated publication volume). We describe and implement both in the following case of study.

Single and accumulated growth indicators can be visualized in order to provide intuitive and actionable insights about the hot topics of a certain domain. Most commonly, trends are visualized by plotting over time as in (Woon et al., 2010). We further seek to integrate the visualization of growth indicators with the underlying knowledge structure (i.e. the taxonomy as provided by the taxonomy creation algorithm of Section "Taxonomy Generation"). To this end we developed two visualization techniques: color coded hierarchies and hierarchical tag clouds. The font size or color of a node reflects the average publication year of the branch under that node. This visualization technique is adapted from tag clouds that are useful as visual information retrieval interfaces (Lohmann et al., 2009). The hierarchical arrangement of tag clouds places semantically related terms nearby which are reported to be advantageous (Hassan-Montero and Herrero-Solana, 2006) for the viewer's perception of the research field. By using font sizes or node colorings to represent growth potential, we are able to overlay this important information on top of the structural information conveyed by the keyword taxonomies. A further important advantage is that encoding the growth indicators in this way allows provides the growth of entire regions in the landscape to be detected – this would be very difficult to do, for example, if the indicators were merely presented as a ranked list of terms.

RESULTS: PROOF OF CONCEPT INSTANTIATIONS APPLIED TO SELECTED CASE STUDIES IN RENEWABLE ENERGY

It is important to note that the system described in Figure 1 is merely a high-level framework, and that each of the five components can be implemented in a variety of different ways. In the following sections we discuss proof-of-concept instantiations in an instructive manner.

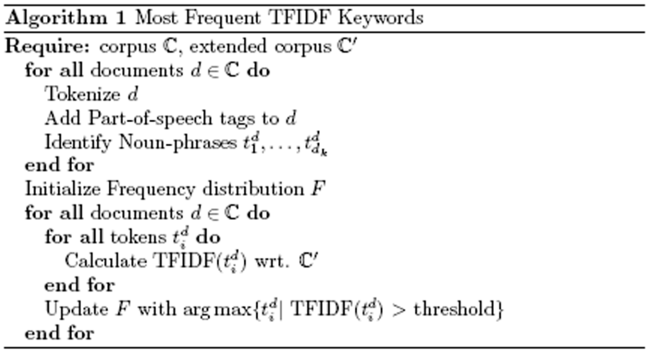
To provide a concrete application domain for developing and testing our tools, it was decided that a case study in a specific domain of technology was required. As our research has been supported by the Masdar Initiative, the natural choice was to conduct a study of the field of renewable energy (RE). It is an active area of research, the ultimate goal of which is to find and exploit new forms of energy; prominent examples include a variety of "green" energy technologies such as wind power, solar heating or biomass. The advancement of RE technologies is of critical importance as traditional resources such as coal, oil and natural gas are finite in nature and are also damaging to the environment. In addition, the research landscape of RE is rapidly evolving and is extremely rich in that research in RE is intricately linked with research in a variety of seemingly unrelated scientific disciplines. The framework outline above provides a decision maker with an eagle eye perspective over a research landscape. Previously, we studied trends in Renewable Energy Desalination and Power Generation (Dawelbait et al, 2010). To further focus our efforts, the examples presented in this chapter will be centered on the research landscape of a particular RE subfields, amongst them biofuel and Waste-to-energy (W2E).

W2E is described as the process in which waste is used to generate electricity or heat. There is a growing interest in this technology since it helps to avoid waste disposal related environmental problems, while serving as a valuable source of renewable fuel. Supported by our web mining methodology, we are interested to undertake an exploratory analysis of the W2E research. Firstly, we will produce a comprehensive hierarchical map which represents the state-of-art of W2E topics, processes, products and infrastructures. Secondly, we are keen to use this map to identify recently growing W2E areas in order to inform the decision making process of science and technology management for possible investments and/or to propose solutions for a sustainable and clean world.

Waste to Energy Publication Corpus

After an initial survey of possible web sources we chose SCOPUS as the central source of data acquisition. SCOPUS is a subscription based,

Figure 2. Algorithm



```
Algorithm 1 Most Frequent TFIDF Keywords
Require: corpus $\mathbb{C}$, extended corpus $\mathbb{C}'$
  for all documents $d \in \mathbb{C}$ do
    Tokenize $d$
    Add Part-of-speech tags to $d$
    Identify Noun-phrases $t_1^d, \ldots, t_{d_k}^d$
  end for
  Initialize Frequency distribution $F$
  for all documents $d \in \mathbb{C}$ do
    for all tokens $t_i^d$ do
      Calculate TFIDF($t_i^d$) wrt. $\mathbb{C}'$
    end for
    Update $F$ with $\arg\max\{t_i^d \mid \text{TFIDF}(t_i^d) > \text{threshold}\}$
  end for
```

professionally curate publication database provided by Elsevier. It provides document records generally of high quality in terms of Meta-data annotation, publication coverage and search term relevance. Assuming that a large selection of relevant documents can be retrieved in a machine readable format, the entire corpus is then preferably stored in a relational database, as in our case study. This solution provides fast and easy access to subsets such as documents from a certain year range. The database also features search enhancing tools such as search indices, full text search and the storage of secondary data such as word-stemmed abstracts and identified noun phrases of abstracts.

Extraction of Waste to Energy Terms

For term extraction we used the most frequent keywords associated to documents by SCOPUS. Additionally we apply the NLP based term extraction algorithm described in Figure 2. The algorithm generates keywords for each abstract by identifying noun phrases and we use a set of tools that are part of the NLTK toolbox (Bird et al., 2009) such as its built-in sentence word Tokenizer.

The identified keywords are scored by the TFIDF scheme [Salton and McGill, 1984], which we refer as TFIDF-keywords. The TFIDF of a keyword t_i is given in Equation 3.

$$TFIDF[t_i^d] = TF[t_i^d] \cdot IDF[t_i] \qquad (3)$$

which is the product of the term frequency $TF[t_i^d]$ (i.e. the number of times a term t_i occurs in a document *d* divided by the number of words in that document) and the inverse document frequency $IDF[t_i]$ (i.e. the logarithm of the number of all documents divided by the number of documents where the term occurs). While the former accounts for the emphasis of a word in terms of repetitions in a document, the latter makes sure that words occurring almost everywhere are downgraded. We then select the most frequent of these TFIDF-keywords.

As a result, the most frequently occurring words of the Waste to energy corpus (shown in Table 1) such as "results", "effect", "study", "paper" are not present in the list of most frequent TFIDF-keywords. Unfortunately, general or more abstract terms that are still useful for taxonomy creation such as "process", "temperature" or "biomass" are also eliminated because they are abundant in the selected corpus. This effect can be mitigated by extending the corpus with unrelated scientific documents (denoted **C'** in the algorithm), such that it will be distinguishable whether terms are specific for a certain research field only or whether they are generally used.

In effect, general research terms such as "experiment" still receive a low inverse document

Table 1. The first column contains the ranked list of the most frequent keywords as annotated by SCOPUS together with its frequencies. The second column list the most frequent, self extracted TFIDF-keywords and their respective frequencies.

Ranking	Keyword	Term Frequency	Keyword	Term Frequency
1.	results	4124	soil	106
2.	study	3562	ethanol	103
3.	biomass	2933	mg/l	98
4.	effect	1876	model	98
5.	paper	1761	hydrogen	90
6.	production	1671	coal	87
7	process	1606	reactors	87
8.	addition	1565	pyrolysis	84
9.	treatment	1520	phenol	83
10.	order	1465	sludge	80
11.	removal	1447	gasification	79
12.	days	1380	sewage sludge	78
13.	waste	1352	biosorption	77
14.	effects	1332	methanol	71
15.	temperature	1325	heavy metals	69
16.	reactor	1298	food waste	67
17.	sludge	1216	reactor	66
18.	presence	1144	biofilm	65
19.	system	1117	biosolids	60
20.	wastewater	1107	membrane	57

frequency, whereas frequent yet domain specific terms such as “biomass” score a higher inverse document frequency and are hence ranked higher. In our case study we therefore offset significant W2E-terms with respect to their occurrences in biomedical literature, extracted from PubMed Central, a freely available database of scientific abstracts. The extracted terms are shown in the rightmost column of Table 1. The final selection is subject to a careful choice of the extended corpus and the involved thresholds (i.e. how many TFIDF-keywords per document and which minimal TFIDF value should be chosen).

Afterwards we applied manual post-processing and stop lists (e.g., geographic terms from SCOPUS’ controlled keyword vocabulary can be useful for classifying publications by geographic location). However, for technology taxonomies these terms are generally irrelevant.

Taxonomy Creation: Fully Automated Taxonomies

Firstly, the fully automated taxonomy creation process was used to analyze the data. As mentioned previously, our approach is based on the Heymann-Algorithm as described in Section “Taxonomy Creation”. To demonstrate the applicability of this approach to technology forecasting, we apply it to a number of different domains. In general, the taxonomies resulting from these analyses are quite large, so what we show in the following pages is

Figure 3. Taxonomy subtree for semiconducting cadmium compounds

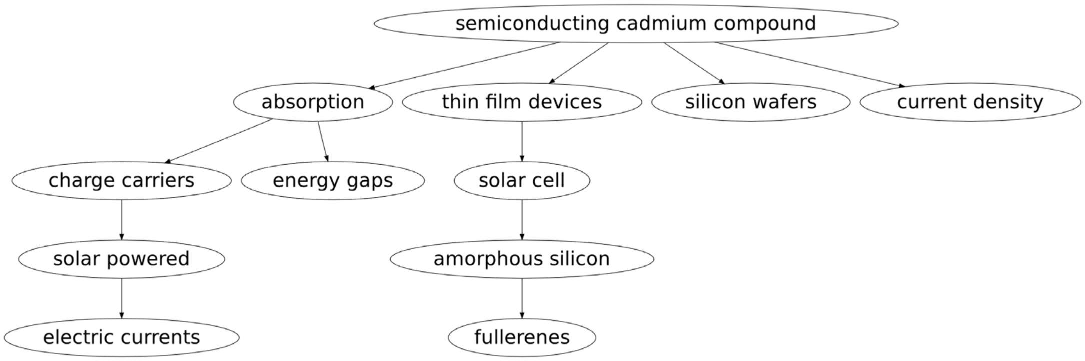
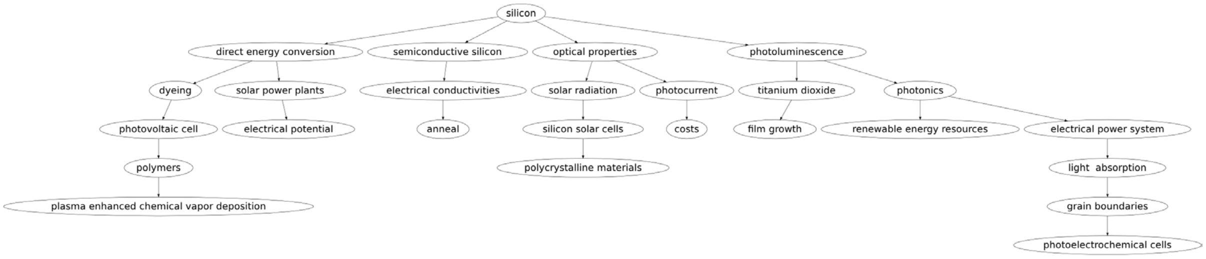


Figure 4. Taxonomy subtree for silicon



a sampling of interesting sub-trees which have been extracted from the corresponding publication corpora.

It is also important to note that the algorithms demonstrated here have a number of variations or "settings" which control the execution of the algorithm. Examples of these include the number and selection of keywords used, the type of centrality measure and the type of similarity metric used to compare the tags. We concede that varying these settings can significantly alter the resulting taxonomy. However it is not within the scope of this chapter to systematically investigate the effect each of these settings has on the taxonomy generation process; instead, readers are referred to (Henschel et al., 2009, Camina, 2010), which provide a much more detailed treatment of this issue. The subtrees presented here are chosen to be typical of the kinds of results that were obtained, and are aimed at providing the reader with an idea of the capabilities of our approach.

Example subtrees from the following three RE-related research domains have been selected and are presented here:

- **Solar PV:** Example subtrees were gen\ erated for the "semiconducting cadmium compounds" (Figure 3) and "silicon" subtrees (Figure 4).
- **Geothermal Energy:** Example subtree for "rocks" was generated (Figure 5).
- **Waste to Energy (W2E):** Example subtrees for this domain were generated for "biomass" (Figure 6) and "wastewater" (Figure 7)

All taxonomies were generated using the Heymann algorithm, and the Sine distance was used to create the distance matrices (this is the distance-based analog of the Cosine distance). The number of keywords used for each taxonomy ranged from 100 to 400.

Figure 5. Taxonomy subtree for rocks

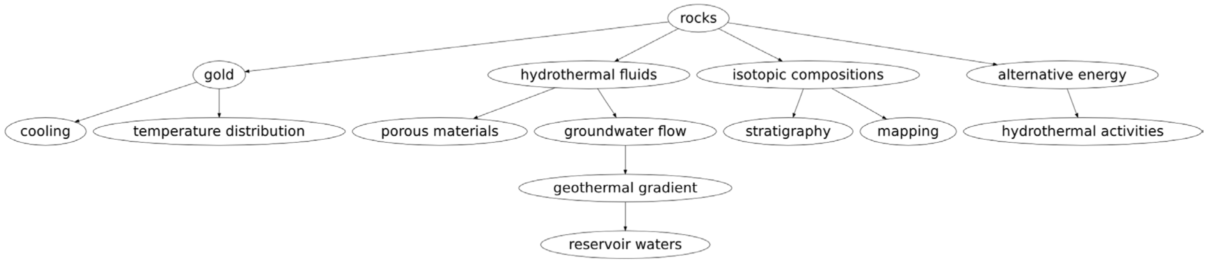


Figure 6. Taxonomy subtree for biomass

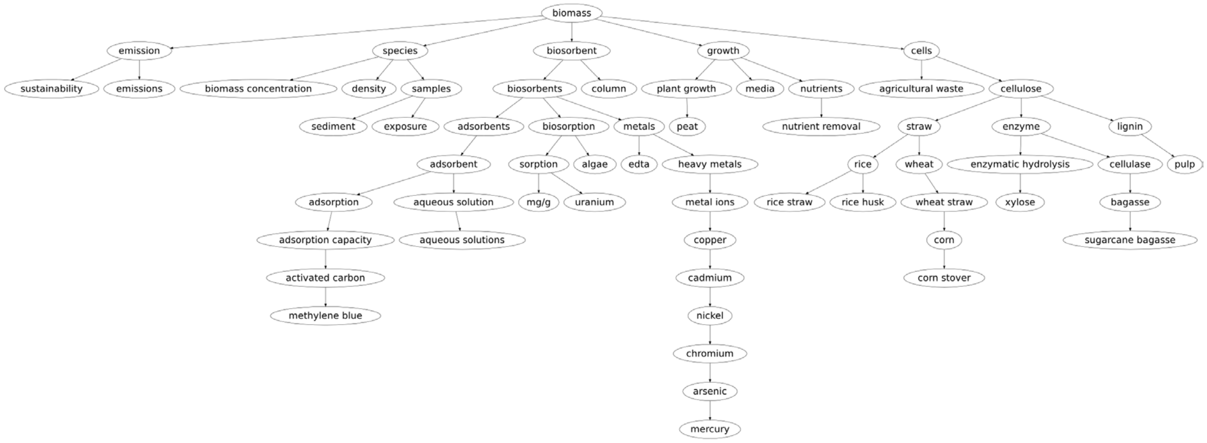
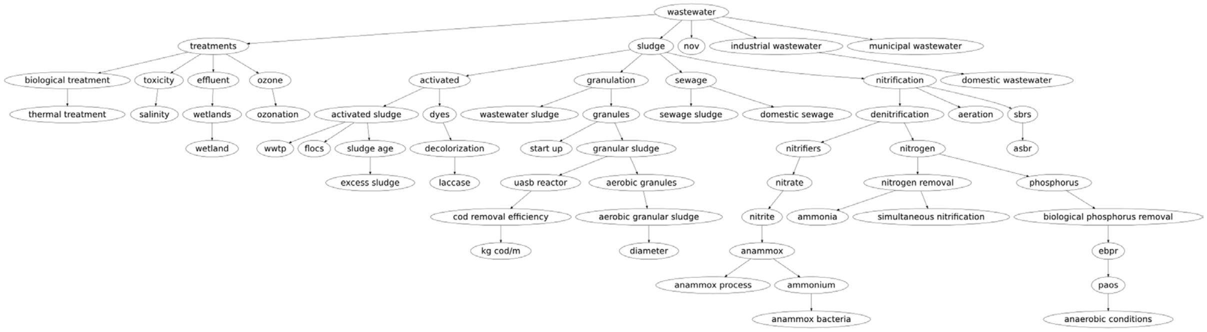


Figure 7. Taxonomy subtree for wastewater



Our main observations are as follows:

1. The quality of the results varied significantly between domains and between subtrees within the same domain.
2. For the "semiconducting cadmium compounds" subtree (Figure 3), it can be seen that a number of related topics were correctly attached, for example *absorption* and the related sub-topics (related to the physics of light absorption), and *thin film devices*, which are an important application of these cadmium compounds.

3. One interesting example where the results are not as good as in "semiconducting cadmium compounds" is in the case of the "rocks" subtree (extracted from the geothermal taxonomy). In this case, we see that the four sub-topics are *gold*, *hydrothermal fluids*, *isotopic combinations* and *alternative energy*. While it is possible to conceive of each of these topics as being related to "rocks", it is quite clear that in this case the algorithm has not produced an informative taxonomic structure. What appears to have happened is that an extremely general term ("rocks"), has, by virtue of its generality, artificially grouped a collection of terms which are otherwise unrelated (or only weakly related).
4. The two W2E subtrees ("biomass" and "wastewater") are significantly larger and more complex than the other subtrees shown here, and helped to highlight the performance of the algorithm with respect to very complex taxonomies. Broadly speaking, the two taxonomies seemed to provide a good illustration of their respective subject areas. However, upon closer inspection, we see that there are a number of irregularities, which would merit further study.
5. In Figure 7, the series of nodes from "heavy metal" to "mercury" represent compounds which are related but which are clearly not subclasses of each other. A similar situation is encountered with the "granulation" to "diameter" path in Figure 8, where we see that each of the three intervening nodes contain some variant of the term "granule". In this example, the similarity function would appear to be picking up semantic relationships rather than actual technological dependencies.
6. As a further example, consider the *thin film devices → solar cells → amorphous silicon fullerenes* branch in Figure 3. On the one hand, these are all topics which are closely related while on the other hand, it is still difficult to explain how *solar cells* came to be a sub-topic of *thin film devices*. We can surmise that the taxonomic structure induced on the keywords might work in a similar fashion to the related technique of hierarchical clustering.
7. These questions raise one of the main problems with the approach, which is that it is difficult to find a clear interpretation of the taxonomic links. While a traditional taxonomy is commonly defined by "is-a" relationships, it is clear that the automatically generated taxonomies do not necessarily follow this rule.

Taxonomy Creation: Incorporation of Expert Knowledge

In conclusion, we note that fully automated taxonomy generation techniques are able to produce results that are interesting. However, it is also faced with a variety of problems. Firstly, automated taxonomy generation is a somewhat inconsistent process which can, under unfavorable conditions, result in inaccurate or noisy results. Secondly, the choice of algorithm settings is also a difficult problem for which there is no straightforward solution.

A viable alternative might be to opt for a semi-automatic process which would allow some prior knowledge to be incorporated into the taxonomy generation process. This allows for the best of both worlds to be enjoyed. On the one hand, we benefit from the advantages of the automatic approach, namely the ability to quickly incorporate the latest developments as well as to efficiently utilize very large quantities of data; on the other hand, taking a semi-automatic approach allows for valuable input from experts and other manually curate sources to be taken into account. By providing a scaffold or framework with which the taxonomies may be initialized, this approach helps to significantly reduce the uncertainty and inconsistency experienced when purely automatic approaches are used.

Depending on the desired accuracy and the final purpose of taxonomies, their fully automated creation remains a very ambitious endeavor. Many researchers have therefore suggested semi-automated protocols, in which experts have manual influence during various stages of the taxonomy/ ontology creation process. The field of Ontology Engineering deals with these aspects. Cimiano points out that automatic extension of existing ontologies have been shown to work successfully (Cimiano, 2006). As a consequence, tools have been created which help to extend ontologies by suggesting terms and their location in the ontology (e.g. within the context of the Gene Ontology project). We therefore investigate the possibility to capitalize on available expert knowledge as an initial guidance to the taxonomy creation process. Note that this approach is an appropriate alternative to the fully automated procedure where expert knowledge is available. We emphasize that we can easily extend the formalism of the Heymann-Algorithm to accommodate initial expert knowledge. The precondition is that terms -at least the expandable nodes- of the expert guidelines must occur frequently in the corpus in order to provide compatibility in terms of the similarity measure. In that case, expert knowledge can be formalized as an initial taxonomy, which is then extensible in the same way, the automatic Heymann algorithm extends a growing taxonomy.

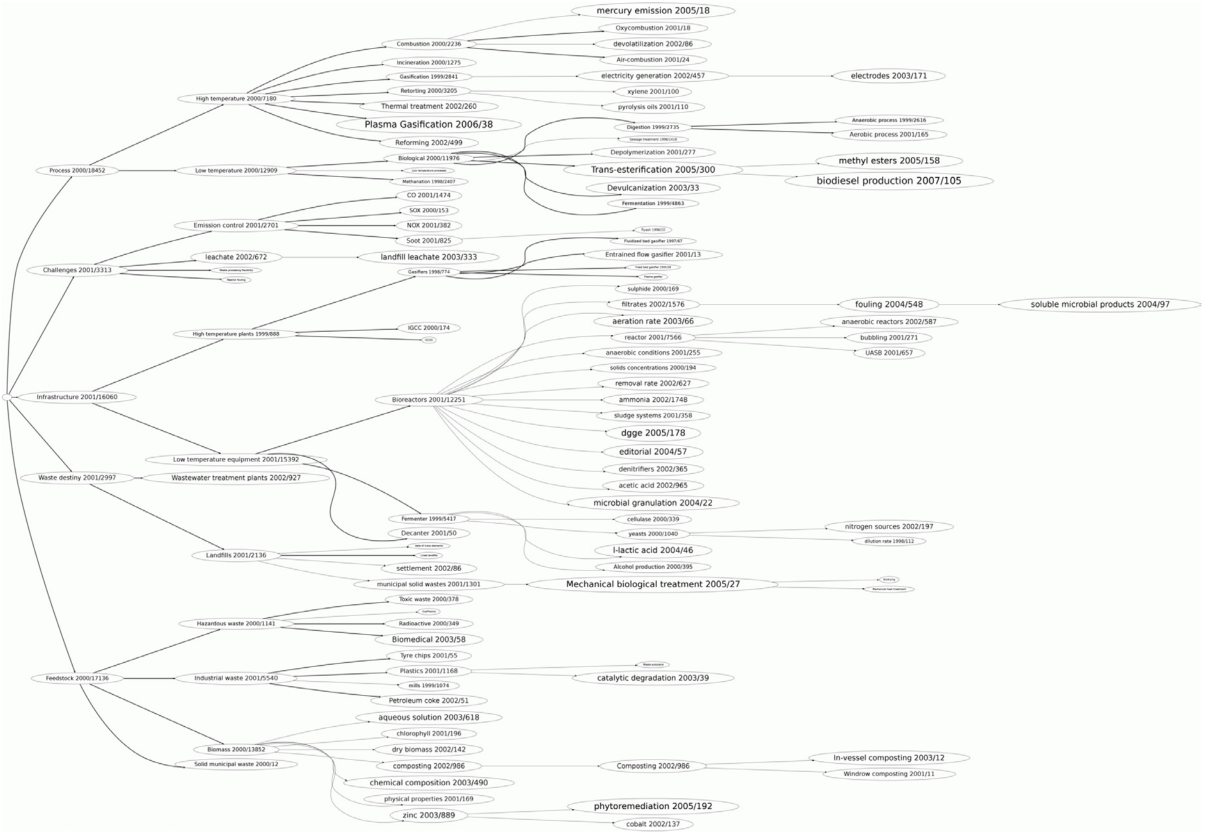
Visualization

Figure 8 shows a taxonomy which has been constructed in collaboration with an expert of W2E technologies. The taxonomy largely consists of predefined taxonomic relations, which are subsequently extended with 100 TFIDF keywords (larger taxonomies can be found in the Supplementary material). In addition, as mentioned in section "Growth indicator accumulation and visualization", the growth indicators were also incorporated into the figure by modulating the font sizes. As mentioned previously, we have used both font sizes and color-codings to convey growth indicators; however using font sizes has an important advantage in that these are preserved when the document is printed in black and white, which is the reason for its use here.

If such a semi-automatically created taxonomy is embedded in the general framework (Figure 1), it is interesting to inspect the growth indicators (i.e. the recency and the volume of the research bodies associated to each node [shown respectively in all nodes]). In particular we note that the font-size modulation allows the growth potential of the nodes to be very clearly visualized. It can for example be seen that the top level categories at taxonomy level 1 and 2 are all balanced out in terms of recency (all are within 1999-2001) due to the average of their associated subtopics. "Plasma Gasification" is the most recent topic (2006). Moreover, it becomes apparent that recently "Biodiesel production" is frequently discussed in the context of "Transesterification". This is evidenced by a body of 105 publications with average publication year 2007. The findings for Biodiesel are consistent with earlier results reported in (Dawelbait et al, 2010) even though a different corpus and a different keyword set was used. The taxonomy created with frequent keywords (Supplementary material S1) unravels that "Removal experiments" have been mentioned in 272 documents with an average date of 2007. A further inspection into the corpus reveals that indeed a large number of recent papers mention different kinds of removal experiments, such as nitrogen removal. Another term that grew to recent popularity is "Wastewater reclamation". We found its mention in 433 papers.

In general, it must be said that the recency of subordinate terms are generally independent from each other (i.e. the W2E research landscape developed rather heterogeneous). This is in contrast to the related study on Renewable Energy (Supplementary material, Figure 2). There, com-

Figure 8. Semi-automatic taxonomy for W2E, incorporating expert knowledge and 100 TFIDF keywords. Average publication year and associated research body is provided for each term. Large fonts indicate strong recent growth.



plete branches including subordinate terms could be identified as hot topics, for example most subordinate terms of "Biofuels".

CONCLUSION

Summary of Findings and Analysis

In this chapter, a novel framework for analyzing and forecasting the growth of technology has been presented. This framework has been developed to efficiently mine online databases in order to enhance R&D operations and inform technological decision-making in a given field of science and technology. The high-level organization of our system comprises a series of computational steps:

1. Data collection from web sources
2. The extraction of domain-relevant keywords
3. The design of growth indicators
4. Automatic and semi-automatic taxonomy generation

In order to validate the technical implementation of our framework we consider possible examples of analysis in the domain of the renewable energy and particularly in the subfield of Waste-to-Energy (W2E). The results of our analysis, as validated by an expert in the field, confirm the benefits of using our approach.

Table 2. A listing of possible future research organized by respective system block

Block	Current approach	Possible future techniques/research directions
Data/ Sources **(A)**	Currently, the Scopus publication database is used.	Future research in this area can be divided into two main topics: • Determination of broader/more diverse sources of publication statistics. Examples include internet-based sources like Google scholar, scirus and Microsoft Academic search, patent databases like Lexis-Nexis, "social" resources like blogs and twitter feeds, technical reports and even the popular press. • Incorporation of more intelligence into the process – a particularly interesting idea would be to place a weighting over the sources depending on the quality or degree of confidence in the database. A simple example would be to use the journal impact factors to determine the weighting of publications appearing in the different databases.
Term extraction **(B)**	Selection of terms is based on term frequencies	A variety of term extraction algorithms have been presented in the literature; these have already been detailed above but briefly, many use more sophisticated methods based on natural language processing (NLP). One notable avenue that we hope to pursue is the extraction of significant or frequently occurring tuples of words (*n*-grams) or noun-phrases. A further direction that could prove important is to involve subject matter experts in the term collection process. This could either be via direct methods (i.e. asking the experts to list interesting terms), or through relevance feedback methods, where the experts is asked to evaluate an automatically generated subset, allowing it to be further fine-tuned.
Growth indicators **(C)**	Average publication year	In other publications, we have already attempted other growth indicators, generally based on estimating the first or second derivative of the publication growth. Avenues for future investigation include: • Identification of more advanced growth indicators, possibly based on technical indicators from the field of finance • Another promising direction is the creation of hybrid indicators that involve the combination of a number of basic growth indicators via committee-based or averaging schemes.
Taxonomy generation **(D)**	In this publication both fully automatic and semi automatic techniques have been explored. However, both approaches are based on variants of the Heyman algorithm.	This is an area that is under intense development and research. The following areas, in particular, have looked promising: • One problem with the Heymann-algorithm is that it is used primarily in batch mode, where a large set of keywords are structured into a taxonomy based on their usage patterns. An alternative strategy would be to use document classification techniques in an iterative manner to partition the research landscape into a hierarchy of categories. The advantage of this approach is that new and previously unseen materials can be quickly added to the existing taxonomy. • Probabilistic topic models are a special class of machine learning methods which represent the content of a document as a mixture of latent variables (i.e. topics) (Steyvers and Griffiths, 2007). A topic is a semantic entity that describes an idea, a concept or an argument the author of a document is expressing through a mixture of words and sentences. In particular, we are keen on using these hidden concepts instead of single keywords as in this chapter to analyze a collection of scientific documents related to the technologies of interest, link their relationships and track their evolution over time. Moreover, since these models are based on a Bayesian framework they can be designed to incorporate prior knowledge to extend and improve the process of taxonomy creation and keyword extraction in the previous stages. • Finally, alternative visualization schemes such as topographic maps are also being considered. These still allow the relationships between domains of research to be easily visualized but, unlike taxonomies, do not impose restrictive and *a priori* structural constraints on these relationships. Such maps can be generated using a variety of techniques such as multidimensional scaling and spring force models.
Growth indicator accumulation **(E)**	Growth indicators for nodes attached to a common ancestor are simply averaged to obtain the overall score.	• Different weighted averaging approaches could be attempted, for example one could devise a scheme that prioritizes more immediate descendants over "distant relatives". • We are also interested in trying more sophisticated approaches that analyze the growth indicators for attached nodes and detects patterns beyond simply aggregate growth. For example, one cluster could contain nodes that are uniformly "high growth", while another might have a mix of extremely high and low growth nodes. While these two example could theoretically have the same average growth rates, there are important qualitative differences between the two which might be important.

However, the results also point to some limitations of a completely automatic process of taxonomy generation, where several inconsistencies in the relationships between keywords have been found. These observations highlight the inherent difficulty of inferring a taxonomic structure from incomplete and noisy observations (such as those obtained from locally cached publication databases). To address this problem, our framework permits the incorporation of prior knowledge (by an expert or by a previously built taxonomy) into the taxonomy generation process, and this was shown to result in more stable results. Such an approach was demonstrated and was able to extract and highlight trends and patterns that were consistent with actual developments in the field of W2E.

Future Directions

As indicated in previous sections, it is important to note that the system described here does not represent the primary value of our work; rather the main innovation is in the overall framework, which describes how a number of separate activities can be combined in a novel way to facilitate the process of technology forecasting. However, the specific selection of algorithms used is by no means optimal (in fact they are intended only as an early demonstration of the potential capabilities of our methodology).

In the context of this formulation, the future development of this work could naturally be organized as the identification of optimal methods for conducting each of the activities depicted in blocks (a)→(d) (Fig. 1). These have been tabulated in the Table 2, as follows.

REFERENCES

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. G. (2007). DBpedia: A nucleus for a Web of open data. In *ISWC* [Berlin, Germany: Springer.]. *Lecture Notes in Computer Science*, *4825*, 722–735.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. Beijing, China: O'Reilly.

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., & Hedlund, T. (2010). Open access to the scientific journal literature: Situation 2009. *PLoS ONE*, *5*(6), e11273.

Camina, S. (2010). *A comparison of taxonomy generation techniques using bibliometric methods: Applied to research strategy formulation*. Meng., Massachusetts Institute of Technology.

Cimiano, P., Pivk, A., Schmidt-Thieme, L., & Staab, S. (2004). Learning taxonomic relations from heterogeneous sources. *Proceedings of the ECAI 2004 Ontology Learning and Population Workshop.*

Cimiano, P., Völker, J., & Studer, R. (2006). *Ontologies on demand: A description of the state-of-the-art, applications, challenges and trends for ontology learning from text*.

Daim, T. U., Rueda, G., Martin, H., & Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, *73*(8), 981–1012.

Dawelbait, G., Mezher, T., Woon, W. L., & Henschel, A. (2010). Taxonomy based trend discovery of renewable energy technologies in desalination and power generation. In *Management of Engineering*. Technology. PICMET.

Frantzi, K., Ananiadou, S., & Mima, H. (2000). Automatic recognition of multi-word terms: the c-value/nc-value method. *International Journal on Digital Libraries*, *3*(2), 115–130.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–901.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Harris, Z. (1968). *Mathematical structures of language*. Wiley.

Hassan-Montero, Y., & Herrero-Solana, V. (2006). *Improving tag-clouds as visual information retrieval interfaces*. In InScit2006: International Conference on Multidisciplinary Information Sciences and Technologies.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the Conference on Computational Linguistics*

Henschel, A., Wächter, T., Woon, W. L., & Madnick, S. (2009). Comparison of generality based algorithm variants for automatic taxonomy generation. *Proceedings of the International Conference on Innovations in Information Technology*, Al Ain, UAE.

Heymann, P., & Garcia-Molina, H. (2006). *Collaborative creation of communal hierarchical taxonomies in social tagging systems*. Technical Report 2006-10, Stanford University.

Hulth, A. (2003). Improved automatic keyword extraction given more linguistic knowledge. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, (pp. 216-223).

Kalfoglou, Y., & Schorlemmer, M. (2003). Ontology mapping: The state of the art. *The Knowledge Engineering Review*, *18*(1), 1–31.

Kostoff, R. N. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*, *68*, 223–253.

Lohmann, S., Ziegler, J., & Tetzlaff, L. (2009). Comparison of tag cloud layouts: Task-related performance and visual exploration. In T. Gross, J. Gulliksen, P. Kotz, L. Oestreicher, P. A. Palanque, R. O. Prates, & M. Winckler (Eds.), *Proceedings of INTERACT, Lecture Notes in Computer Science, 5726*, (pp. 392–404). Springer.

Maedche, A., & Staab, S. (2001). Ontology learning for the Semantic Web. *IEEE Intelligent Systems*, 16.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Martino, J. P. (2003). A review of selected recent advances in technological forecasting. *Technological Forecasting and Social Change*, *70*(8), 719–733.

Miller, G. A., Beckwith, R., Fellbaum, C. D., Gross, D., & Miller, K. (1990). WordNet: An online lexical database. *International Journal of Lexicography*, *3*(4), 235–244.

Navigli, R., Velardi, P., & Gangemi, A. (2003). Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, *18*(1), 22–31.

Noy, N. F. (2004). Semantic integration: A survey of ontology-based approaches. *SIGMOD Record*, *33*(4), 65–70.

Pazienza, M., Pennacchiotti, M., & Zanzotto, F. (2005). *Terminology extraction: An analysis of linguistic and statistical approaches* (pp. 255–279). Berlin/ Heidelberg, Germany: Springer.

Porter, A. (2005). Tech mining. *Competitive Intelligence Magazine*, *8*(1), 30–36.

Porter, A. (2007). How tech mining can enhance R&D management. *Research Technology Management*, *50*(2), 15–20.

Salton, G., & McGill, M. (1984). *Introduction to modern information retrieval*. McGraw-Hill Book Company.

Sanderson, M., & Croft, B. W. (1999). *Deriving concept hierarchies from text*. In ACM SIGMIR Conference on Research and Development in Information Retrieval, (pp. 206–213).

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web revisited. *Intelligent Systems*, *21*(3), 96–101.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. In Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.), *Handbook of latent semantic analysis*. Hillsdale, NJ: Erlbaum.

Wächter, T., & Schroeder, M. (2010). Semi-automated ontology generation within OBO-Edit*Bioinformatics, 26(10)*. Oxford University Press.

Wang, P., Hu, J., Zeng, H. J., & Chen, Z. (2009). Using Wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, *19*(3), 265–281.

Witten, I., Paynte, G., Frank, E., Gutwin, C., & Nevill-Manning, C. (1999). KEA: Practical automatic keyphrase extraction. *In Proceedings of the 4th ACM Conference on Digital Library.*

Woon, W. L., Henschel, A., & Madnick, S. (2009). A framework for technology forecasting and visualization. In *Proceedings of the International Conference on Innovations in Information Technology*, Al Ain, UAE.

Woon, W. L., & Madnick, S. (2009). Asymmetric information distances for automated taxonomy construction. *Knowledge and Information Systems*, *21*(1), 91–111.

Woon, W. L., Zeineldin, H., & Madnick, S. (2010). (in press). Bibliometric analysis of distributed generation. *Technological Forecasting and Social Change*.

ENDNOTES

[1] http://www.nactem.ac.uk/software/termine

[2] http://developer.yahoo.com/yql/